AMERICAN ASSOCIATION OF WINE ECONOMISTS

AAWE WORKING PAPER
No. 239
Economics

IF THIS WINE GOT 96 OUT OF 100 POINTS, WHAT IS WRONG WITH ME? A CRITIQUE OF WINE RATINGS AS PSYCHOPHYSICAL SCALING

Denton Marks

May 2019

www.wine-economics.org

If This Wine Got 96 Out of 100 Points, What Is Wrong with Me? A Critique of Wine Ratings as Psychophysical Scaling

Denton Marks
Professor
Department of Economics
University of Wisconsin-Whitewater, USA
Email: marksd@uww.edu

If This Wine Got 96 Out of 100 Points, What Is Wrong with Me? A Critique of Wine Ratings as Psychophysical Scaling

Denton Marks

Department of Economics, University of Wisconsin-Whitewater, USA

(marksd@uww.edu)

"We cannot share experiences so we cannot compare perceived sensations directly."
(Bartoshuk et al., Physiology and Behavior 82 (2004), p. 110)

Abstract

Along with technological changes, dispassionate expert wine evaluation to educate consumers might seem aimed at increasing market efficiency since consumer ignorance likely inhibits wine market growth, despite increased global affluence and a growing middle class. Considerable research has explored how ratings correlate with transaction prices, testing whether they help "explain" willingness to pay (WTP); that could suggest that they do indeed educate consumers.

While mixed results say that wine ratings are not necessarily reliable guides to wine quality and WTP, a more fundamental structural difficulty is that, as a form of hedonic quality index, they involve questionable interpersonal comparisons—say, between experts or between an expert and oneself. For example, different tasters may differ over a taste's appeal (e.g., the existence of "supertasters"). Saying that experts can tell consumers what they will like so they can reliably determine relative enjoyment and willingness to pay is flawed logic.

This paper explores wine ratings as a form of hedonic psychophysical scaling, using the scaling literature as its analytic framework. Well-established critiques of hedonic scaling illuminate the difficulties and raise fundamental questions about the reliability of ratings—in effect, adopting someone else's preferences as one's own—and the interpretation of any price-rating correlation.

Food scientists have developed techniques to address the challenge of interpersonal hedonic evaluation (e.g., "magnitude matching" which attempts to normalize hedonic ratings against a putative common yardstick). While this might lead to more reliable expert ratings, the underlying point is that interpersonal comparisons are unreliable absent shared preferences. Use of wine ratings in econometric work of fine wine price determinants—and by *the public more generally in evaluating relative enjoyment of wines—has been largely uncritical of their underlying logic. The logical problems are as much psychophysical as econometric; and, probably because of their interdisciplinary nature, the existing literature seems to have overlooked their importance.*

Key words: fine (premium) wine, experts, ratings, psychophysical scaling, hedonic index

*The College of Business and Economics, University of Wisconsin-Whitewater, has provided financial support for this research. I presented a related paper at the 10th Annual Conference of the American Association of Wine Economists (AAWE), Bordeaux (2016), where I received a number of helpful comments. Victor Ginsburgh has also provided useful guidance. Benoit Lecat (California Polytechnic State University), Nick Vink (Stellenbosch University), and Steve Charters (Burgundy School of Business/School of Wine and Spirits Business) have provided both helpful and stimulating discussions and generous hospitality during sabbatical year (2018-19) academic visits that allowed me to continue this research.

INTRODUCTION

The consumer's problem in forming a willingness to pay (WTP) for wine, especially fine wine, is one of the thorniest in wine market analysis. The problem raises good questions from the philosophical—What does it mean to know something? How can we know a wine and communicate such knowledge?—to the practical—Is the wine market smaller because of this problem? If so, how much? Much of the problem involves the value and validity of wine ratings.

Because of the consumer's problem, the content of wine economics—relative to viticulture or oenology or other fields that study wine—has given particular prominence to the role of experts and their ratings. Ashenfelter has characterized the role of experts as one of the two central questions in wine economics (2016), and one can imagine that it is at least as important in wine economics as in other areas of cultural economics (e.g., Ginsburgh 2016). Among the most popular themes in wine economics has been the determination of fine wine prices. Among the most prominent questions in that literature is the impact of expert ratings, although the evidence of a relationship is mixed and limited to a narrow, though financially important, segment of the wine market—results from commercial auctions of classified Bordeaux (e.g., Oczkowski and Doucouliagos, 2014; Luxen, 2018) which provide rare published data on "market-clearing" transactions. Despite that limitation, the attention ratings have received even in that context is good reason to scrutinize what is behind them.

PSYCHOPHYSICAL SCALING: RATINGS AS HEDONIC SCALES

Food processors use various techniques to determine the characteristics of their products and the extent to which consumers find them appealing. Arguably, many important techniques of this type are psychophysical scaling (PS): "…the process of quantifying mental events [responses], especially sensations and perceptions, after which it is possible to determine how these quantitative measures of mental life are related to quantitative measures of the physical stimuli [signals]." (Marks and Gescheider, 2002, p. 91) PS includes many techniques to measure how much consumers like a product and the intensity with which they experience it. In the broadest sense of the phrase, winemaking is food processing.

Simply put, wine ratings resemble ratings on a hedonic scale. A hedonic scale is a "numerical scale used to indicate degree of liking and/or disliking." (Lawless, 2013, p. 391) The hedonic scale is familiar as the worst-to-best scale we use in describing preferences or, if one is familiar with sensory research, the scale that registers one's degree of like or dislike of examples of some good (e.g., brands of orange juice).

An expert's ratings of wines within some "reference group" are simply scores along some spectrum of like to dislike with (a) the extremes defined by the expert's opinion of the best and worst of the reference group and (b) tasting notes provided to support the rating. Table 1 provides a representative example of a hedonic scale. Given the pattern of wine ratings, one might need to redefine the extremes or, alternatively, one could agree that experts discard a grade or two at the bottom and work with only six or seven with the lowest perhaps "Dislike moderately" (value = 3). For example, Robert Parker equates ratings in the 96-100 range of his 50-to-100 point scale as "an extraordinary wine of profound and complex character" (his emphases). He does not explain the meaning of 100 points—wine vendors are known to label such wines with the troubling term "perfect"—but "Like extremely" is a good estimate. At the other end, Parker gives wines 50-59 points if they are "unacceptable". A reasonable inference is that he would "dislike extremely" a 50-point wine.

The variety of ratings available is wide, but all fall within this general pattern. Table 2 illustrates how one might align Parker ratings with the hedonic scale given in Table 1.

ISSUES AND OUR FOCUS: INTER-INDIVIDUAL COMPARISONS

A large literature critiques PS. Reviews of the literature have appeared (e.g., Lim 2011), and even a summary of the major issues would take us far beyond the scope of this paper. However, we can highlight several prominent issues relevant to wine ratings as hedonic scales:

- The environment in which the scaling occurs, uncontrolled influences on subjects (external such as "pollutants", internal such as the impact of emotion and memory), and mismeasurement of controlled influences (e.g., recent wine tasting experiences).

Table 1: A Standard 9-point Hedonic Scale

9-Point Hedonic Scale	
9	Like extremely
8	Like very much
7	Like moderately
6	Like slightly
5	Neither like nor dislike
4	Dislike slightly
3	Dislike moderately
2	Dislike very much
1	Dislike extremely

Table 2: Wine Ratings as Hedonic Scale: Example (signal=flavor; response=rating)

Comparison				
9-Point Hedonic			Wine Ratings (Parker)	
9	Like extremely		96-100	Extraordinary
8	Like very much		90-95	Outstanding
7	Like moderately		80-89	Very good to barely > average
6	Like slightly			
5	Neither like nor dislike		70-79	Average, soundly made, little distinction
4	Dislike slightly			
3	Dislike moderately		60-69	Below average, notable deficiencies
2	Dislike very much			
1	Dislike extremely		50-59	Unacceptable

- Respondent difficulties with the measurement scale—numeric, spatial, verbal, number and size of intervals available. Do respondents find that the scale used allows them to provide an accurate representation of the relative degree of liking and disliking?
- Intra-individual consistency (the extent to which an individual gives consistent responses to the same stimulus): this is particularly important in wine rating since an expert should give the same rating to the wine stimulus in subsequent evaluations. Evidence of intra-individual consistency with hedonic scaling generally seems nonexistent because of the difficulty in performing careful subsequent tests with the same respondents. Evidence from carefully conducted subsequent blind wine tastings is at least as scarce.

Of particular importance to this discussion is the additional issue of inter-individual comparisons—in effect, what consumers do when they formulate their WTP using one or more expert (hedonic) wine ratings, comparing the expert's scaling to their own scaling or perhaps even adopting it as their own.

Unlike intra-individual differences, inter-individual differences in scaling have received considerable attention in the PS literature. Two prominent researchers state flatly:

> "Magnitude estimation functions of individuals [(MEF): the cognitive provision of a numeric response to the stimulus] vary substantially….An important problem in psychophysical scaling has been to determine how much of this variability reflects real interindividual variation in the relation between stimulus and sensation….[M]agnitude estimates given by individual observers cannot be meaningfully compared in any simple or direct manner. " (Marks and Gescheider, 2002, pp. 120, 123)

For our purposes, Lim (2011, p. 739) has two key criticisms of the hedonic scale in his survey. First:

"…due to its inequality of scale intervals and the lack of a zero point…the scale can yield only ordinal- or, at best, interval data (i.e., ordered metric). Thus, the scale cannot provide information about ratios of liking/disliking for stimuli…nor provide meaningful comparisons of hedonic perception between individuals and groups…"

That is, first, the steps on the hedonic scale provide simply an ordering of levels of preference. For example, in Table 1, we do not know if the sensory distance between ratings of 6 and 7 equals the sensory distance between ratings of 7 and 8 (e.g., as modifiers to "liking", the difference between "slightly" and "moderately" need not be the same as the distance between "moderately" and "very much"). If these were cardinal measures, then we could say that the two distances are equal (one "liking" unit or perhaps, in utility terms, the same number of "utils") and make accurate statements about percentage increases or decreases in enjoyment.

By the same logic, one cannot know that a one-point change in a wine rating always means the same thing—for example, the one-point difference that could represent the change from "Extraordinary" to "Outstanding" and then "Outstanding" to "Very Good" in Table 2. In contrast, we know the meaning of a one-dollar change in price, a cardinal measure, regardless of the point from which the change occurs. Second, the lack of a common zero point means that "the ratio of the numbers assigned to objects has no meaning": for example 50 degrees of temperature is more thermal energy than 25 degrees, but we cannot say that it is 100 percent warmer or twice as warm (Cardello, 1998, p. 14). An 96-point wine is better than an 80-point wine, but we cannot conclude that it is 20 percent better. We do not know the meaning of a 0-point wine—or even a 50-point wine, which sets the lower bound in Parker's scale.

For our purposes, Lim's second important point is that:

> "…from a statistical standpoint, because the data it yields are categorical and discrete without a true zero point, the type of statistical analyses that can be applied with confidence is limited, i.e., nonparametric statistics. However, it is common practice for researchers to use more powerful parametric statistics, such as analysis of variance, to analyze data collected with the scale, although it is mathematically inappropriate to do so."

That is, we cannot assume that we know the distribution of hedonic data and the parameters of the distribution so it is inappropriate to preform statistical analyses that assume otherwise such as multivariate regression. Thus, one cannot treat hedonic ratings like cardinal numbers like weight or price.

Bartoshuk is among the most persistent and incisive critics of misusing hedonic ratings, noting not only the misapplication but also persistent failure to stop it:

> "The deep problem is that, except for fictional mind readers, people cannot share each other's experiences of pleasure and pain. Yet with the misuse of [hedonic] scales, we

> sometimes act as if we can….Incidentally, this error keeps being rediscovered." (2014, p. 91)

It is understandable that behavioral scientists would like to be able to make valid interpersonal comparisons and apparently continue to do so, but the desire and the effort do not mean that we have the tools to do it.

This is the issue. If we agree that ratings are hedonic scales and observe that individuals' MEFs vary substantially, then we must ask how one can, in effect, reliably predict own enjoyment by adopting someone else's MEF as one's own. How is it reasonable for the consumer to assume that the expert's MEF mimics her own? Bartoshuk et al. (2004: p. 110) state the problem clearly: "We cannot share experiences so we cannot compare perceived sensations directly."

An additional complication of applying PS to wine is wine's complexity. PS focuses upon the evaluation of a single sensation; among flavors, it might be saltiness. Hedonic evaluation of foods often tests the relative appeal of versions of the same food with the versions differing only by the amount of some ingredient (e.g., sugar): one can isolate how liking varies with the amount of the ingredient. There is a reference point relative to which the analyst scores the alternatives.

Wine rating lacks such a reference point. A comparison of wines may specify a peer group as the basis for a rating (e.g., '12 Margaux), but the reference group is rarely identified precisely and is far less precisely defined than a given food product being tested with variations in one ingredient.

MAGNITUDE MATCHING: "INTER-" COMPARISONS

Research on problems with hedonic scales illuminates some known issues with ratings. Sensory researchers encountered a problem when they applied traditional 9-point hedonic ratings to questions such as taste preferences between men and women or, more generally, "across-subject" or interpersonal comparisons—the kind of comparison consumers often want to do in comparing expert ratings or, more importantly, in comparing an expert's differences with their own. Subjects have different frames of reference for sensory evaluation, and subjects indicating the same difference in liking might actually be experiencing very different changes in liking and vice versa.

A recent paper illustrates the problem as well as a prospect for improved methodology for making "inter-" comparisons. Kalva et al. (2014) focus upon the criticism of inappropriate inter-individual comparisons: "…on occasion, these [hedonic] scales have been used to make across-subject/group comparisons. This is a very different psychophysical challenge….These errors result because we cannot directly compare sensory or hedonic experiences." (2014, pp. S238-S239).

Here is their approach. Give Group I (GI) subjects two 9-point scales on which to rate a variety of foods: an hedonic scale that measures liking or disliking (1=Dislike extremely to 9=Like extremely) and a sensory scale that measures the intensity of stimulation (1=No sensation of the taste (e.g., bitterness) to 9=Extreme sensation). Ask a statistically comparable (e.g., gender, age) Group II (GII) of subjects, first, to identify four extreme experiences in their lives to set boundaries:

- Among <u>sensory</u> experiences, "no sensation" (sensory value = 0) and the "strongest imaginable sensation of any kind" (sensory value = 100); and
- Among <u>hedonic</u> experiences, the "strongest imaginable liking of any kind" (hedonic value=+100) and "strongest imaginable disliking of any kind" (hedonic value=-100).

Given that the object of the GII rating exercise involves food, the boundary setting must not involve food items: thus, for the sensory standard, they come from other sensory experiences such as loudness, brightness, or pain.

We interpret the resulting range of intensity to be the same for each GII member—not the same source of the intensity standard but the same "zero to most" range. When these subjects are given a food to taste (e.g., black coffee), those with the more sensory-sensitive palates (e.g., "supertasters") will give a higher intensity rating. If we test a number of foods, we can assign a sensory intensity score (SIS) for taste to each subject—say, from 0 for no taste to 10 for the most sensitive supertaster.

Why can we not allow food to be the basis for the sensory intensity standard? Assume that we do and that we know somehow that half the subjects are normal tasters (say, SIS = 5) and the other half are supertasters (say, SIS =10). When we have them all taste black coffee, they may all rate it as 70 out of 100. However, the taste intensity is not actually the same: the intensity to the supertasters would be much greater than that of the normal tasters on a sensory scale that is independent of taste. [1]

Similarly, we also want the hedonic scale for GII subjects to be independent of food.

In effect, these most-least scales for GII subjects—known as general Labeled Magnitude Scales (gLMS)—set the absolute boundaries of their lifelong sensory and hedonic experiences. They can differ between any two GII subjects—we have no objectively absolute standards for measuring hedonic and sensory reactions—but each GII subject is still comparing the foods to "the same" subjective absolute scale, namely, the limits of each's hedonic and sensory experience. While the non-food sensations specified can differ, the fact that they are all comparing against most and least extreme experiences means that, in that sense, they are using the same absolute scale, and comparing their relative rankings is closer to an absolute standard with this than without. The formal name for this technique is the method of "magnitude matching" (p. S239)—that is, establishing matching scales against which we measure individual differences in sensory and hedonic reactions.

The idea is that we norm the ultimate like-dislike scale to the subjects' best and worst experiences ever: "Similar to the sensory gLMS, the key property of the hedonic gLMS is that it assesses liking for a particular stimulus (for example, food) in the context of all affective experience." (Ibid., p. S239) We are getting closer to the same absolute scale for all which, if known, would allow us to take anyone's scale and compare it to anyone else's: A has the same intervals as B but is twice as wide; the top half of C's scale is the same as the bottom half of D's scale, and so forth. Assume that there is in nature an absolute scale of sensory sensitivity (pain, brightness, loudness, taste intensity, etc.) with a zero point minimum like the Kelvin scale for temperature. For a given sensation, everyone has a range of sensitivity along that scale, but it is possible that no two scales match in change sensitivity, breadth, or otherwise.

Sensory differences like this resemble the sensory differences with other mammals. Some have greater acuity or sensitivity to changes in stimuli (e.g., visual detection of movement), some have a wider range of sensitivity in one direction or both (e.g., hearing), and some have less sensitivity (e.g., color changes and differences). We know also of differences in taste sensitivity among species—comparative hedonics (e.g., Beauchamp and Mason, 2014).

Arguably, this method provides more meaningful interpersonal comparison. In their experimental setting, Kalva et al. found that subjects with different taste sensitivities (sensory) exhibited different levels of liking and disliking (hedonic)—in particular, statistically significant positive correlations with liking and negative correlations with disliking. A number of subsequent studies have acknowledged the problem addressed by magnitude matching and have used gLMS (e.g. Williams et al., 2016).

In the experiment, ask the GI subjects to rate one or more foods on each of the 9-point scales (e.g., how much s/he likes black coffee and the extremity of the taste sensation). Ask the GII subjects to rate the same foods using a gLMS which ranges from -100 to +100 for the hedonic (dislike-like) scale and from 0 to 100 on the sensory (no sensation-greatest possible sensation) scale.

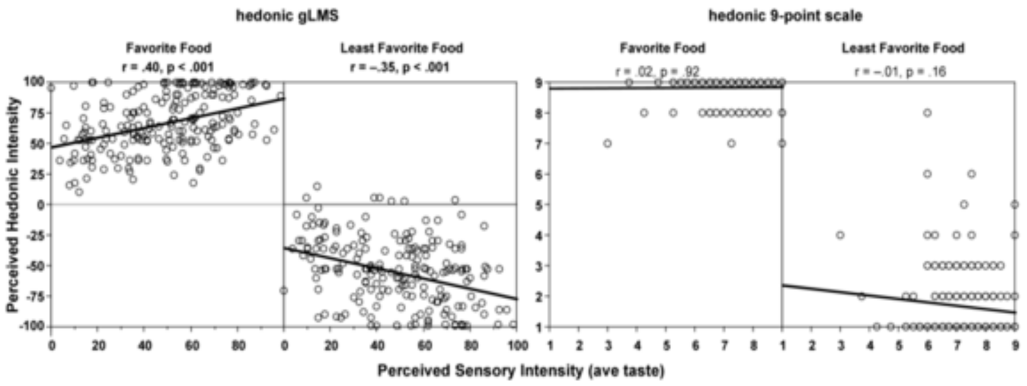
The study's results demonstrate that those using the gLMS consistently exhibit a strong correlation between hedonic intensity of liking (positive) or disliking (negative) and perceived (sensory) intensity of the sensation. For example, the more intensely one experienced a favorite food, the higher the hedonic rating in the context of imagining the hedonic and sensory extremes one has ever known. The authors attribute the differences in sensory intensity to the presence of different degrees of "tasters"—minimal, medium, and super, for example. For those using simply the 9-point scales, the correlations between liking/disliking and perceived sensory intensity were much weaker and almost never statistically significant. Figure 1 provides an example of this. Kalva et al. explain:

> …the hedonic 9-point scale cannot "see" this effect [of individual differences in taste intensity] as correlations between food hedonics and taste [sensation] are not significant. Although the extreme labels on the hedonic 9-point scale ("like extremely" and "dislike extremely") do not explicitly refer to food, asking subjects to rate their favorite and least favorite foods reveals that, in fact, subjects do treat the top of the hedonic 9-point scale as maximum food palatability [sensation] and the bottom as minimum food palatability [sensation]. This is not surprising. In the absence of other instructions, subjects reasonably treat the scale labels as referring to the subject of the test. In other words, the gLMS shows that maximum and minimum food palatability [sensation] vary with individual differences in taste perception, but the hedonic 9-point scale obscures this effect because it implicitly assumes that the extremes of food palatability are equally intense for everyone." (p. S241; emphasis added)

While this research model does not produce valid interpersonal comparisons generally (e.g., utility), it highlights a fundamental problem in comparing two or more hedonic ratings and provides one technique for addressing it.

The significant correlations it found between taste sensitivity and degree of liking or disliking are, for our purposes, less important than its demonstration that "we cannot compare perceived sensations directly". While the hedonic differences from the analysis with conventional hedonic and sensory scales are not significant, those with the hedonic gLMS certainly are. We are learning more about the relationship between differences in liking and disliking and differences in other characteristics such as taste sensitivity.

Comparison of the Hedonic General Labeled Magnitude Scale with the Hedonic 9-Point Scale



Journal of Food Science
Volume 79, Issue 2, pages S238-S245, 14 JAN 2014 DOI: 10.1111/1750-3841.12342
http://onlinelibrary.wiley.com/doi/10.1111/1750-3841.12342/full#jfds12342-fig-0003

Figure 1: gLMS and 9-Pt Hedonic Comparison (used with permission)

What is the relevance for wine ratings? Acting as if we can simply compare two conventional hedonic ratings to see how similar or different two coffee drinkers are—or taking one's rating (the expert's) as a good predictor of another's (the consumer's)—is the mistake that use of the gLMS attempts to correct. Acting as if we can simply compare two conventional hedonic ratings to see how similar or different two wine tasters are—or taking one's rating (the expert's) as a good predictor of another's (the consumer's)—is, at best, naïve.

Consider Figures 2 and 3. Figure 2 depicts individuals who are all identical in all characteristics relevant to evaluating wine flavor--the assumption required for the inter-individual comparison between expert and consumer to be reliable. If we are all alike, then our hedonic responses will duplicate the rater's responses, *cet. par*., and we would all give the wine the same rating. If one or more persons does not know the wine, then they can simply adopt the rating of those who do. If we can take the content of much direct-to-consumer wine marketing copy as an indicator, many wine consumers use that model or something like it.

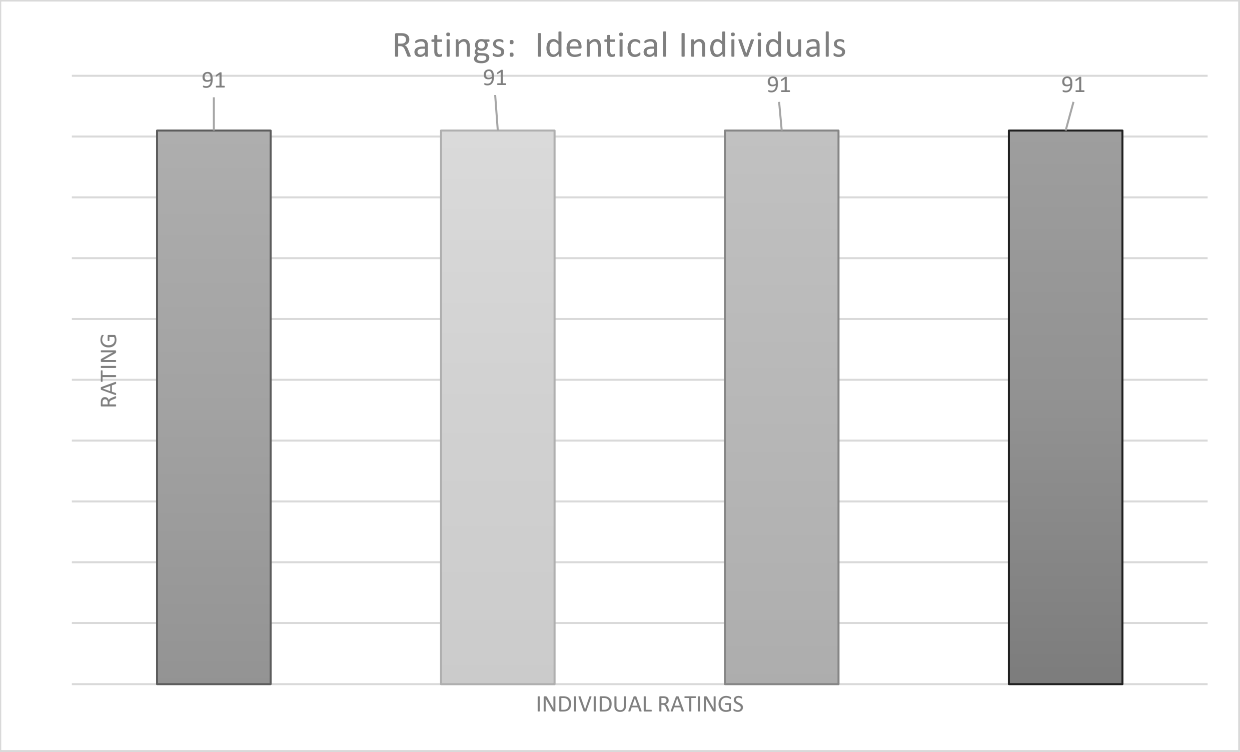


Figure 2: Identical Individual Sensory Sensitivity

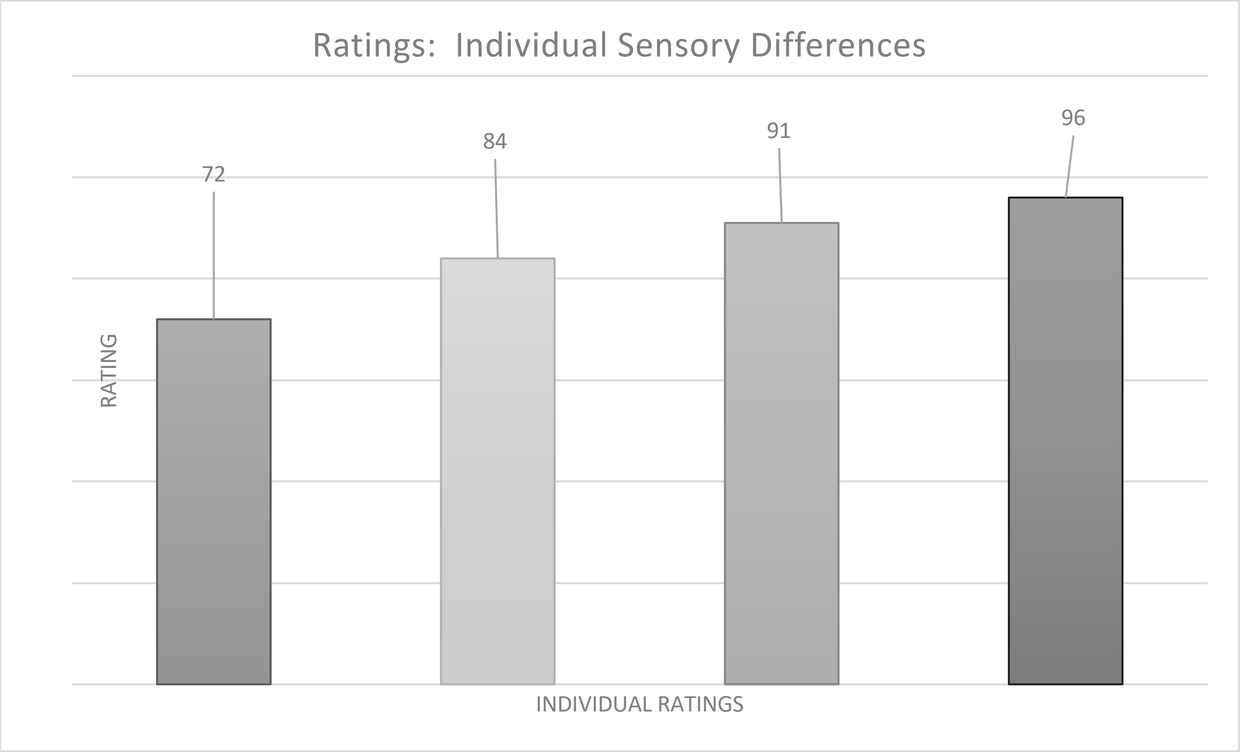


Figure 3: Increasing Individual Sensory Sensitivity

Figure 3 reflects individual differences along some measurable difference in individual characteristics—for example, sensitivity toward a pleasant taste, which is a sensory measure, not hedonic. If we are all different and we sort individuals by that distinguishing characteristic using a common (non-food) scale for imagining like/dislike (hedonic) limits, then we might see more sensitive tasters deriving relatively more pleasure (as Figure 1 also suggests). Figure 3 illustrates how those with different sensitivities experience different degrees of enjoyment and, of course, resembles Figure 1.

We can generalize Figure 3 at least to any characteristic to which we could apply sensory intensity. The psychophysical literature often mentions pain, loudness (hearing), and brightness (vision); but we can imagine others—anything that involves sensory perception from hearing, seeing, smelling, tasting, touching, or other less familiar senses (perhaps as many as 33 (e.g., Blakemore 2014)). Beyond that, we can imagine other differences that might matter (e.g., gender, ethnicity, age), although the source of those differences might, in fact, be differences in senses that correlate with those characteristics.

This illuminates several persistent problems with wine ratings. We can start with the first concern of Bartoshuk as one of the pioneers in the psychophysical research: "we fail to see differences that are real" (2014, p. 92). If the expert likes it and the consumer does not, what is wrong? Perhaps or probably, nothing is wrong: finding one or more or all expert ratings unhelpful may be fully rational and healthy. The focus of the psychophysical research is, in part, finding ways to compare individuals because we cannot simply "compare perceived sensations directly"; we are different.

Second, if wine ratings are hedonic ratings, then it highlights our ignorance of the reference points of the experts. It is as if the experts are using the traditional 9-point scale; and, in order to assume comparability among their ratings, we must treat them as all having the same palate. Notwithstanding all of the other difficulties inherent in interpreting a rating (e.g., Marks, 2015), this is the confusion that consumers experience regularly when they find more than one rating of a wine, but the ratings conflict: "…one consistent finding that remains indisputable is the known vast differences in the rating of wine quality by putative wine experts" according to Cicchetti and Cicchetti (C/C) (2014, p. 34). They conclude that those seeking guidance seek out experts whose palates seem to correspond with their own and follow them (Ibid.). Their results substantiate the noncomparability of individual ratings and confidently open the door to questioning the usefulness of ratings as a reliable guide for forming WTP on the basis of expected own enjoyment.

Once we move from robust reliability to conditional reliability—the consumer needs to find the expert that best reflects her personal palate—then we seem relegated to a crude trial-and-error process of seeking only the most general guidance to wine enjoyment. Given the variety of sources of ratings beyond individuals—especially tasting groups or panels and their changing composition or the delegation of ratings from the expert to the expert's staff member—then the search for a fellow traveler becomes more complicated.

This also questions the meaning of the evidence of correlation between ratings and wine prices. Ashenfelter and Jones (2013) provide evidence from the most heavily studied price-ratings relationship—numerous studies of various ratings sources compared to Bordeaux auction prices—that suggests that those prices may correlate positively with ratings because buyers value the ratings *per se*, not because they add information about the drinkability of the wine. Like evaluations of postage stamps and fine art, "…it is also possible that the experts' ratings influence prices because they create values that are independent of the function [of the wine—drinkability] and thus become self-fulfilling prophecies." (p. 293)

Third, given the C/C results, one must question the comparability of ratings and the conversion of widely disparate rating systems (50-100 points, 0-3 stars, etc.) into one common metric—lacking an alternative, the technique C/C use in their study (e.g., multiply ratings on a 10-20 scale by 5 so they can be compared to a 50-100 scale). Can we know reliably that 3 out of 4 on a 4-point scale is reliably equivalent to 87.5 on a 50-100 scale or 75 out of 100 on a 1-100 scale?

Fourth, experts provide ratings relative to some often poorly defined reference or peer group—an intervening variable that typically does not appear in gLMS rating and which differs from the global extremes of liking and sensation used by gLMS. The lack of a well-defined reference group further clouds interpretation and manipulation of the group definition can elevate or diminish a wine, depending upon one's objectives.

Finally, given the difficulties just described, it should not be surprising that finding clear and consistent correlations between ratings and wine prices (assuming that we are looking at actual transaction prices and not asking prices) has eluded us. Trying to do that would resemble taking someone's ratings of orange juice and analyzing the extent to which they correlate with orange juice prices. Given the range of consumers' affinity for orange juice; and, among those who like it, the reasonable differences of opinions that emerge, it would be surprising if any expert's ratings had any particular influence upon WTP—unless enough consumers defer to the expert and say "If she likes it, then I should like it, and that determines my WTP".

Whether one could conduct a comparable study of wine quality or whether magnitude matching is exactly what is missing in interpreting and comparing expert ratings is less the point than the concept that such ratings do not allow one to know the individual differences among the experts. If one's palate happens to align well with an expert's palate who also happens to taste many of the wines that one considers, then that seems fortunate. However, the frequency of that happening is likely small relative to the number of consumers trying to know what they are doing when they shop for fine wine.

OUR INTEREST IN EXPERTS

Our focus has been the scientific support for consumers' tendency to use expert wine ratings to anticipate their own enjoyment of a wine—in particular, whether the sensory reaction of an expert to a wine is a reliable guide to or forecast of one's own reaction. A more fundamental question is why we turn to experts. In an important book—a "major intellectual event" according to one prominent economist (Schleifer, 2012: p. 1)—Kahneman (2011) discusses "Intuitions vs. Formulas" or the choices we make for guidance in evaluation and forecasting between expert intuition and algorithmic formulae such as regression-based forecasting models. For his example of a formula that offers "a compelling demonstration of the power of simple statistics to outdo world renowned experts" (p. 223), he happens to choose Ashenfelter's Bordeaux wine price model (e.g., Ashenfelter, 2008). Perhaps ironically, he proceeds later in the chapter to criticize the unnecessary complexity of "multiple regression" relative to simpler, equal-weight linear relationships (p. 226).

Kahneman's discussion serves at least two purposes. First, he discusses the psychology of preferring one approach to the other; and, second, he discusses the circumstances under which one or the other might be more reliable.

Part of his focus is our desire for validity. If we seek evidence, we are more likely to be persuaded by coherence, familiarity, credentials, and/or quantification. Most of us want the world to "make sense", and that is the basis for most of our biases and heuristics: favor those who are familiar or look successful, assume that higher price means higher quality, and so forth. Trust others' experience, even if we cannot evaluate success or record.

He posits (pp. 228-9) that the bias toward human intuition reflects a "deep resistance to the demystification of expertise". It comes from (a) favoring humans over machines when they compete and (b) the human preference for the natural over the synthetic or artificial, citing the example of increased WTP for organic foods, despite no evidence of greater benefit. Generally, we are more accepting of errors from humans than from algorithms, especially as the consequences increase. He leaves us wondering about the origins of these preferences.

The long lists of adjectives that Kahneman draws from Meehl's pioneering work in comparing clinical predictions with statistical ones in psychology (1954) convey the flavor of the bias. Two of the most evocative from the 20 applied to statistical methods are "sterile" and "blind". Two of the most evocative from the 20 applied to clinical (intuitive) judgment are "holistic" and "living" (p. 228).

Kahneman does not deny the existence of expertise but has arrived at two conditions that vary directly with the development of expertise: (a) the regularity of the environment and (b) the time taken to learn the regularities of the environment. Since he also discusses this, we can take the second condition to include a timely feedback mechanism that reinforces correct learning. Thus, one is more likely to become an expert by spending more time in a more regular environment with more timely feedback and less likely to become an expert otherwise. At the other extreme are "low validity" environments less likely to nurture expertise.

Applying this to wine quality requires that we think more carefully about the content of wine ratings. They have at least two dimensions: the quality of the product and whether the consumer will enjoy it. The environment of the former is relatively regular and is the subject of lifelong study by serious students of wine: traditional and permissible varietals, classic blends, impact of environmental conditions such as weather and soil, winemaking techniques (e.g., use of wood, types of yeasts, fermentation techniques), and so forth. It is relatively uncontroversial that some can become experts at rating the quality of a product relative to its caricature expression.

The other dimension of rating is predictions of enjoyment. This is a problem. Slipping from an evaluation of product quality to prediction of enjoyment usually happens—for example, when experts use descriptors like "sexy" or even the more modest "delicious" and especially when they encourage purchase ("incredible value") and particularly volume purchase ("buy this by the case", "back up the truck"). How can a stranger advise that reliably? This may occur for several reasons, but one of them is likely that this is what consumers seek: they defer to experts and desire validity. Most are unlikely to know enough to evaluate the validity of a technical judgment; what influences purchasing is advice about enjoyment. Experts are inclined to accommodate this; but then they have slipped from technical evaluation to hedonic liking/disliking evaluation, and the concerns raised in the preceding analysis demonstrate why that is problematic.

CONCLUSION

Wine consumers who venture beyond commodity wines and wines they know well often look to experts and their ratings and tasting notes to guide their WTP and purchase. If personal enjoyment is a theme in the expert evaluations, then consumers may, in effect, be adopting the hedonic preferences of one or more experts in the process. We have examined lessons from the psychophysical literature that questioned simple comparisons of subjects liking and disliking and recognized that individual hedonic preferences are different and that we cannot make such naïve comparisons. Bartoshuk has pushed further and emphasized that this is a persistent problem: we want to make such comparisons, the problems are subtle, and it is tempting to forge ahead (for an example of research that acknowledges the problem but chooses to ignore it rather than re-do the data collection, see Pickering and Hayes, 2017). If we are not careful, we will continue to make the same mistake.

Magnitude matching is a method of moving those who are registering hedonic preferences closer to a shared scale—"the context of all affective experience". Our focus has been explaining the similarity of hedonic scales to wine rating scales and extrapolating from the critique in the psychophysical literature to what we see wine consumers doing with expert ratings. Research on magnitude matching has demonstrated the problem with simple comparisons by illuminating individual differences that such comparisons mask: "we fail to see differences that are real". This has been more effective in illuminating the problem than in offering a complete solution. Some form of magnitude matching might someday facilitate more reliable use of expert ratings, but that is a distant hope.

We have not taken the analysis beyond a recognition that an assumption that another's hedonic evaluation of wine is a reliable to one's own hedonic evaluation of wine is naïve. If it seems that the literature relating expert ratings to transaction prices provides evidence to the contrary, then we can think of a number of reasons why such a correlation can exist without concluding that consumers pay more because they agree with experts who say it is better wine:

- Because of the importance of examining transaction prices, virtually all of the evidence is from commercial auctions of red Bordeaux wine where a number of motives beyond personal enjoyment of the bidder may drive the bidding. These might include potential resale in a wine investment market tuned to ratings where ratings, not flavor, make wine collectible; and bidding by restaurants who must use ratings for marketing the wine and winning wine list awards. Unlike individual enjoyment, ratings are the only, albeit flawed, common metric between commercial transactors so commercial bidders are relegated to paying more for higher ratings. Because the identities of bidders in such auctions are confidential, the studies of price-rating correlations have never identified what share of successful bids come from commercial bidders as opposed to consumers. Beyond that, among consumers as bidders, we cannot know how much they are valuing the rating and not the wine *per se* as suggested by Ashenfelter and Jones.

- Studies have not been careful about adjusting for possible simultaneity between the rating and the price: many experts consider price in their ratings and are rating quality per dollar, not simply quality. When this is true, price and rating are determined simultaneously, and failure to adjust econometrically for this means that the measured effect of rating upon transaction price may be biased upward.
- The relationship between ratings and prices is certainly not reliable and essentially an unknown outside this investment sector of the wine market because of the absence of tests for the partial correlation between ratings and market-clearing prices.

The discussion closes by discussing the psychology of why we turn to experts in the first place—less because they provide answers and more because we have nothing better. We conclude that wine experts have a role, but it is evaluating the technical quality of wine, not how much one might enjoy it. Therein lies the source of the problem.

ENDNOTES

1. Bartoshuk offers the following example (2014). Ask a group of otherwise equal, normal tasters (GN) to rate the sweetness of a soda and ask a group of otherwise equal supertasters (GS) to do the same thing. Assume that both groups rate the sweetness 7 on a 10-point sweetness scale. It looks like they are experiencing the same sweetness. However, perform the same exercise, but give all GN and GS subjects headphones and have them set a tone to a loudness level that corresponds to the sweetness they experience. We assume that (a) unlike taste, both groups have the same sensitivity to loudness, and (b) loudness and taste sensitivities are independent.

Say that the GN group sets it to 80 decibels and the GS group sets it to 90 decibels—twice as loud: "the soda tastes twice as sweet to [the supertasters]" (p. 92)

BIBLIOGRAPHY

Ashenfelter, O. (2008), Predicting the quality and prices of Bordeaux wine", *The Economic Journal*, Vol. 118 No. 529, pp. F174-F184.

Ashenfelter, O. (2016). Remarks at the Plenary Session, American Association of Wine Economists (AAWE), 10th Annual Conference, Bordeaux FR, 21-25 June.

Ashenfelter, O. and Jones, G. (2013), The demand for expert opinion: Bordeaux wine", *Journal of Wine Economics*, Vol. 8 No. 3, pp. 285-293.

Bartoshuk, L., V. Duffy, B. Green, H. Hoffman, C-W Ko, L. Lucchina, L. Marks, D. Snyder, and J. Weiffenbach (2004), "Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching", *Physiology and Behavior,* Vol. 82 No. 1, pp. 109-114.

Bartoshuk, L. (2014), "The measurement of pleasure and pain", *Perspectives on Psychological Science*, Vol. 9 No. 22, pp. 91-93.

Beauchamp, G and Mason, J. (2014), "Comparative hedonics of taste" in Bolles, R. (Ed.), *The Hedonics of Taste*, Psychology Press, New York and London, pp. 159-184.

Blakemore, C. (2014), Rethinking the senses: uniting the philosophy and neuroscience of perception, available at: https://ahrc.ukri.org/documents/case-studies/rethinking-the-senses-uniting-the-philosophy-and-neuroscience-of-perception/ (accessed 18 February 2019)

Cardello, A. (1998), "Perception of food quality" in Taub, I. and Singh, R. (Eds.), *Food Storage Stability*, CRC Press, Boca Raton LA, pp. 1-38.

Cicchetti, D and Cicchetti, A. (2014), "Two enological titans rate the 2009 Bordeaux wine", *Wine Economics and Policy,* Vol 3 No. 1, pp. 28-36.

Ginsburgh, V. (2016), "On judging art and wine" in Rizzo, I. and Towse, R. (Eds.), *The Artful Economist: A New Look at Cultural Economics,* Springer, New York NY, pp. 245-265.

Kahneman, D. (2011), *Thinking Fast and Slow*. Farrar, Strauss, and Giroux, New York.

Kalva, J., C. Sims, L. Puentes, D. Snyder, and L. Bartoshuk (2014), "Comparison of the hedonic general labeled magnitude scale with the hedonic 9-point scale", *Journal of Food Science,* Vol. 79 No. 2, pp. S238-S245.

Lawless, H. (2013), *Quantitative Sensory Analysis*, Wiley-Blackwell, Somerset UK. ProQuest ebrary.

Lim, J. (2011). "Hedonic scaling: a review of methods and theory," *Food Quality and Preference,* Vol 22 No. 8, pp. 733-747.

Luxen, M. (2018), "Consensus between ratings of red Bordeaux wines by prominent critics and correlations with prices 2004–2010 and 2011–2016: Ashton revisited and expanded", *Journal of Wine Economics,* Vol. 13 No. 1, pp. 83-91.

Marks, D. (2015), "Seeking the veritas about the vino: fine wine ratings as wine knowledge", *Journal of Wine Research,* Vol. 26 No. 4, pp. 319-335.

Marks, L. and Gescheider, G. (2002). "Psychophysical Scaling," in Pashler, H. (Edition Ed.) and Wixted, J. (Volume Ed.), Stevens' Handbook of Experimental Psychology, Vol. 4 (3d Edition), Wiley, New Jersey, pp. 91-138.

Meehl, P. (1954), *Clinical vs. Statistical Prediction: A Theoretical Analysis and Review of the Evidence*. University of Minnesota Press, Minneapolis.

Oczkowski, E. and Doucouliagos, H. (2015), Wine prices and quality ratings: a meta-regression analysis. *American Journal of Agricultural Economics*, Vol. 97 No. 1, pp. 103-121.

Pickering, G. and Hayes, J. (2017), "Influence of biological, experiential, and psychological factors in wine preference segmentation", *Australian Journal of Grape Wine Research*, Vol. 23 No. 2, pp. 154-161.

Schleifer, A. (2012), "Psychologists at the gate: a review of Daniel Kahneman's *Thinking, Fast and Slow*", *Journal of Economic Literature*, Vol. 50 No. 4, pp. 1-12.

Williams, J., Bartoshuk, L., Fillingim, R., and Dotson, C. (2016), "Exploring ethnic differences in taste perception", *Chemical Senses*, Vol. 41 No. 5, pp. 449-456.